FACULTY OF SCIENCE AND TECHNOLOGY

MASTER'S THESIS

Study Programme/Specialisation: Mathematics and Physics/Mathematics	Spring semester, 2021 Open

Author: Bharat Ghimire

Supervisor: Tore Selland Kleppe

Title of master's thesis: Towards Bayesian Estimation of Compartmental Models for COVID-19

Credits (ECTS): 60

Keywords: Bayesian statistics,
Markov chains,
MCMC algorithms, Compartmental Models,
Parameter estimation in SIR model.

Number of pages: 74

Preface

I would like to begin by thanking my supervisor Prof. Dr Tore Selland Kleppe for his continued guidance and support throughout my thesis. I would also like to express my sincere gratitude for providing me a book (Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin, 3rd edition) whose ideas and concepts I greatly benefited and drew on to write this thesis.

Abstract

This thesis explores concept of compartmental modelling of infectious disease particularly like COVID-19. The world is still under siege by COVID-19. I have discussed two such fundamental mathematical models SIR and SEIR model that form the basis in epidemiological modelling and how they help predict the dynamics of current pandemic, COVID-19. My research is principally based upon SIR model and 'rstan' codes developed by (Grinsztajn, Semenova, Margossian, & Riou, 2020). The main aim of my thesis is to show how MCMC algorithm is used to estimate the model (SIR) parameter and draw much needed inference about the dynamics of such infectious disease. For this purpose, we tried to use real COVID-19 data from various cities within Norway and started off accordingly. However, the model we used i.e. SIR model did not work perfectly with the data as MCMC algorithm could not sample from huge population size (5.3m) and our sample being too small. We had some glitch in our inference table as clearly evidenced by much variation in the value of $\hat{R}$ implying our Markov chains are not in sync with one another. So, we decided to show the same estimation process with different data. While my focus is on SIR and SEIR model, I have also tried to give the basic concept behind Bayesian thinking and some important MCMC algorithms that are an integral part of such models. I have tried not to digress from COVID-19 to other diseases. The main objective of all the infectious models that statisticians

currently use or will use in future is to predict the dynamics of such disease so that public health officials make informed decisions to save as many lives as they can. I am trying to present how such process works. As more complex mathematical models are being developed around the globe, it is still very useful to understand the SIR and SEIR model as most of the newly developed models are derivatives of these models.

Table of Contents **Page number**

Introduction to Bayesian Statistics:

Thomas Bayes (1701 - 1761)

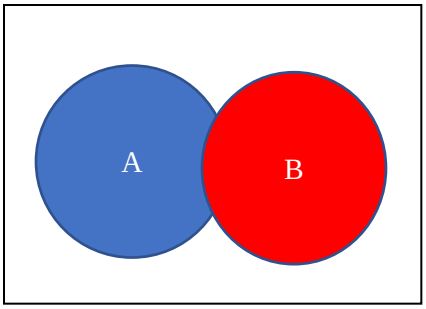
The Bayesian statistics is named after a famous English philosopher and statistician Thomas Bayes (1701-1761). However, the picture shown here may or may not be his. There is no other portrait of him that is available as of today. He first came up with the theorem in his papers named "an essay towards solving a problem in the doctrine of chances" (Bolstad & Curran, 2016). His notes, after this death, would lay the foundation of new approach to statistics called Bayesian statistics or Bayesian inference as we call it today. To understand Bayesian approach, we begin by understanding Bayes theorem.

Probability theory is nothing, but common sense reduced to calculation.

Pierre-Simon Laplace (1749-1827).

Bayes Theorem: Bayes theorem or sometimes Bayes rule is a mathematical formula that premises upon the concept of conditional probability. A conditional probability of an event is the probability of occurrence of that event given a different event has already happened.

Let us assume A and B are two events given shown in the figure below with $P(B) \neq 0$.



(1)

Then mathematically Bayes theorem can be written as following.

$$P(A/B) = \frac{P(B/A)\,P(A)}{P(B)} \tag{2}$$

Where,

P(A/B) = Probability of happening event A given B has already happened or Posterior belief and is given by,

$$P(A/B) = \frac{P(A \cap B)}{P(B)} \tag{3}$$

P(A) = Probability of happening event A or Prior belief.

P(B/A) = Probability of event happening B given the event A has already happened or Likelihood.

And P(B) = Probability of happening event B or normalization factor.

If we have two events A_1 and A_2 then above formula can be modified as,

$$\boxed{\frac{P(A_1/B)}{P(A_2/B)} = \frac{P(A_1)}{P(A_2)}\,\frac{P(B/A_1)}{P(B/A_2)}} \tag{4}$$

Where,

$\frac{P(A_1/B)}{P(A_2/B)} = Posterior\ odds,\ \frac{P(A_1)}{P(A_2)} = Prior\ odds\ ,\ \frac{P(B/A_1)}{P(B/A_2)} = Bayes\ factor.$

We can further extend this if we partition sample space of A as (Beichl, Sullivan, & Engineering). Our new posterior belief can be rewritten as following.

$$P(A_i/B) = \frac{P(B/A_i)\,P(A_i)}{\sum_j P\,(B/A_j)P(A_j)} \qquad (5)$$

Here we see that Bayes theorem is all about application of conditional probabilities. We update our prior belief based on new evidence available to us. For instance, if we want to know the probability of selling an umbrella in a sunny day, Bayes theorem tells us the probability of selling umbrella on other days, rainy windy and snowy. I would like to further mention one of the many applications of Bayes theorem, the breast cancer treatment. Using Bayes statistics, we can quantify the probability of having breast cancer given the person is male or female.

Probability of being a female given he has breast cancer is given by,

P(female/breast cancer) = $\frac{P(breastcancer/female)\,P(female)}{P(breast\ cancer)}$

Thanks to the government record we can have the data of probability of having a breast cancer given the person is female and overall probability of having breast cancer for any individual. Using these information we can update our prior beliefs to arrive at posterior belief. i.e. probability of being a female given the individual has a breast cancer. Bayes theorem or statistics is widely used in financial markets to assess risk, in medicine to test the efficacy of new drugs, and in many other biological experiments. It has given rise to a new field in statistics called Bayesian inference in contrast to frequentist inference or classical statistics. Let's dive into Bayesian inference before moving to Bayesian estimation methods.

Bayesian Inference: The Bayesian inference technique is a statistical method to draw inference based on Bayes theorem. As described above it gives us the so called 'posterior probability' using two entities prior probability and likelihood function clearly demonstrated in following formula.

$$P(H/E) = \frac{P(E/H).P(H)}{P(E)}$$

Where,

H→ Hypothesis whose probability is to be determined.

E → Evidence pertaining to the data which was not used to calculate prior probability.

P(H) → Prior probability

P(H/E) → posterior probability

P(E/H) → The likelihood function

P(E) → Marginal likelihood or sometimes model evidence.

In the light of a new evidence, we make a more informed decision. This decision is thus associated with maximum value of posterior probability often called MAP. For instance, in case of disease diagnosis process, the Bayes rules can be rewritten as,

$$P(disease|symptoms) = \frac{P(symptoms|disease)\,P(disease)}{P(symptoms)}$$

Which follows, $Posterior = \frac{Likelihood.Prior}{Marginal\ Likelihood}$

Bayesian Inference for Probability Distribution:

To present it mathematically how it works, let's assume 'x' as observed data point and 'Θ' be our parameter of our distribution. Bayesian statistics assumes the probabilities for both data point and hypothesis. Θ is a variable and we assume we have prior distribution of the hypothesis i.e. P(Θ), likelihood of the data point P (Data | Θ) or P(x | Θ). The posterior distribution also commonly known as the distribution of the parameters after considering the observed data points is the essence of Bayesian inference. Unlike point estimate, using Bayes theorem, we find the whole posterior distribution of the parameters as described by following equation.

$$P(\Theta|data) = \frac{P(data|\Theta)\,P(\Theta)}{P(data)} \quad (6)$$

Or,

$$P(\Theta|x) = \frac{P(x|\Theta)\,P(\Theta)}{P(x)} = \frac{P(x|\Theta)\,P(\Theta)}{\int P(x|\Theta)\,P(\Theta)\,d\Theta} \quad (7)$$

Where P(Θ|x) is the posterior distribution of the parameter Θ.

$P(x|\Theta)\ P(\Theta)$ is called sampling density for the data.

$P(\Theta)$ is called prior distribution of the parameter.

$\int P(x|\Theta)\ P(\Theta)\ d\Theta$ is called marginal probability of the data. It can be seen here we multiplied sampling density by prior density and integration is taken over sample space Θ. This quantity sometimes also referred as 'marginal likelihood' and works a normalizing constant. It simply scales the posterior density to make it proper density.

It is sometimes also expressed as a proportional relation as posterior α likelihood.prior distribution.

$$\text{Posterior } \alpha \text{ Likelihood. Prior} \tag{8}$$

The posterior distribution is the distribution of new data point, marginalized over posterior.

$$p(\tilde{x}|X) = \int p(\tilde{x}|\Theta)\ p(\Theta|X)\ d\Theta$$

The prior data point is the distribution of new data point marginalized over prior.

$$p(\tilde{x}) = \int p(\tilde{x}|\Theta)\ p(\Theta)\ d\Theta$$

Where, $\tilde{x}$ is a new data point whose distribution we want to predict.

The main aim of Bayesian inference technique is to predict the distribution of new, unobserved data point. In contrast to the frequentist statistics where we try to find optimal point estimate of the parameters by e.g. MLE, then we put the estimated value into a formula in order to find the

distribution of new data point Bayesian approach takes into account the uncertainty in the parameters and therefore estimating the variance of the new predictive distribution correctly.

In proportionality relation above, posterior density is proportional to the likelihood function of the data multiplied by prior of parameters.

The prior distribution in most cases represents is normalized and therefore represents the true density for the parameter. The likelihood function, on the other hand, is in fact a product of many densities and hence does not have normalizing constant to be a true density function. Potential densities function, mostly, need to have a normalizing constant to be a proper density function. However, we need to bear in mind that normalizing constant does not and cannot change the relative frequencies of the random variables. Rather, it only scales up/down the density function.

Let's take an example of normal distribution to get more clarification.

Prior: $\Theta \sim N(0, 1)$

Likelihood: $y|x, \Theta \sim N(\Theta x, \sigma^2)$

Now, with the help of Bayes rule

Posterior distribution is given by

$$\textbf{posterior: } P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{\int P(Data|\theta)P(\theta)\,d\theta} = \frac{P(y|x,\theta)P(\theta)}{\int P(y|x,\theta)P(\theta)\,d\theta} = \frac{(\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-\theta x)^2}{2\sigma^2})(\frac{1}{\sqrt{2\pi}}\exp(-\frac{\theta^2}{2})}{\int(\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(y-\theta x)^2}{2\sigma^2})(\frac{1}{\sqrt{2\pi}}\exp(-\frac{\theta^2}{2})\,d\theta}$$

In order to predict posterior predictive distribution, we use posterior distribution and likelihood distribution

prediction (posterior predictive distribution):

$$P(y^*|x^*, Data) = \int P(y^*|x^*, \theta)P(\theta|Data)\, d\theta = E_{\theta \sim P(\theta|Data)}[P(y^*|x^*, \theta)]$$

Summary of the Bayesian Technique: The Bayesian estimation technique can be summarized in following steps.

i) Formulating the knowledge about problem to be solved.

ii) Collecting data

iii) Getting posterior data i.e. updated beliefs

First step can be achieved in following way.

a) Defining model of distribution: The very 1st step is to come up with a model that entails qualitative as well as quantitative aspects of our problem. The model is thus now associated with unknown parameters which we call random variables.

b) Formulating a prior distribution: The prior distribution entails the uncertainty in our unknown parameters before we have the data.

c) Calculating posterior distribution: In final step we calculate posterior probability distribution with the help of observed data and by estimating unknown parameters. This gives us a new set of updated beliefs.

The posterior probability distribution gives us all the information about parameter Θ. The fact that posterior distribution entails the information from data points, it is thus less variable than prior distribution as described in following equation

$$Var(\Theta) = E(Var(\Theta|x) + Var(E(\Theta|x)) \quad (9)$$

The equation above tells us that the expectation of posterior variation is smaller than that of prior variation by an amount depending on the value of posterior means over the distribution of possible data. The greater the prior variation, the more chances for reduction our uncertainty regarding Θ. The mean and variance relations only tell us about the expectations and in particular situations the posterior variance can be similar to or even larger than the prior variance (although this can be an indication of conflict and inconsistencies between the sampling model and prior distribution.

Significance in parameter estimation: The most fascinating aspect of Bayesian inference technique is that we do not have lots of data. Even with a small number of observations, we can update our prior beliefs. In other words, it facilitates us to update our beliefs continuously as new data comes in. The whole estimation technique follows the same steps that I described above. As we are provided with more and more new data, our posterior belief becomes prior. Then new prior is updated with the help of likelihood derived from the new data and we get a new posterior distribution. This process continues indefinitely as long as we new data comes in and we can continuously update our beliefs.

It is very convenient to employ what you know so far with whatever data available and predict what future would look like. Bayesian inference technique exactly does so. It has long been used in various kinds of sciences mainly because of following reasons.

1) It entails the concept of confidence.
2) It performs well even with sparse data.
3) Model and results obtained are easy to understand and interpret.

The technique allows us to incorporate new opinions and domain-specific knowledge into account. These beliefs are combined with data to constrain the details of the model. Then, we make a prediction based on that model. The model does not give a single answer instead it gives us a distribution of possible answers giving us room to assess uncertainty. The Bayesian estimation techniques have led us to new world of statistical computations that helped revolutionize several disciplines in science.

In real life scenarios, the method has been frequently used in drug testing, monitoring, disease diagnosis, infectious disease modelling, machine learning techniques/algorithms, artificial intelligence and so forth. It has become an integral part of modern statistics and machine learning techniques. My thesis will focus on one of many applications of Bayesian inference technique which is infectious disease modelling.

Markov chain Monte Carlo Method:

The shift from frequentist statistics to Bayesian statistics over the time led mathematician around the world to look for algorithms that could help solve high dimensional probability distribution as well as complex mathematical integrals. Scientists were needing algorithms for solving numerical approximations of multidimensional integrals in different disciplines like, particle physics, computational physics, computational biology, computational statistics and so forth. Valuable contribution from mathematician 'Andrey Markov' helped achieve one of the greatest algorithms of 20th century -MCMC. One recent research puts MCMC algorithm among the top ten algorithms of 20th century that has greatest impact on science and engineering today (Beichl et al., 2000). Over the last two recent decades, it has played a monumental role in shaping not only statistics but in physics, computational science and econometrics (Andrieu, De Freitas, Doucet, & Jordan, 2003).

In statistical computing Markov chain Monte Carlo Method abbreviated as MCMC is a set of algorithms used to sample from a high dimensional probability distribution when direct sampling is difficult. Unlike in Monte Carlo sampling method, which enables us to draw independent samples from the distribution, MCMC allows us to draw next samples that are only dependent on the current samples by developing a Markov chain. Some of the famous MCMC algorithms being

Gibbs sampling and Metropolis-Hastings. To understand how it works in real, let's begin by understanding first what Markov chain is.

Markov Chain: A Markov chain, named after a Russian mathematician Andrey Markov, is defined as a sequence or chain of events in which the probability of each new event depends solely on the previous state. Mathematically, the two types of Markov Chain that we study are as following.

1) Discrete-time Markov chain: It is a chain of random variables X_1, X_2, X_3,…with Markov property or sometimes called memoryless where the probability of landing on the next step depends only current state. It is denoted as $X = (X_n : n \geq 0)$ or $X = (X_n)$ where n = (0,1, 2, 3…). Here the random variables take the discrete values such as integers and hence the stochastic process is called discrete valued. The total possible collection of values that X_n can take is called state space

2) Continuous-time Markov chain: A continuous-time Markov chain $(X_t)_{t>0}$ is a state space with some transition matrix 'A' with dimension equal to that of state space and initial probability distribution defined on that specific space. A_{ij} is non-negative for all $i \neq j$. It

basically, tells us the rate of transition from one state i to another state j. Some of the famous examples Markov chain are random walk, Brownian motion, fluctuation in stock market curve, foraging trace of animals and so forth.

In its basic sense, MCMC was developed by combining two methods Monte Carlo and Markov chain so that it allows us to sample randomly from a high dimensional probability distribution that

premises upon the probabilistic dependence among samples by creating a Markov chain which entails Monte Carlo sample.

MCMC method in Bayesian inference and posterior distribution analysis: From Bayes theorem we get what it requires to calculate posterior distribution. They are a prior distribution, an evidence and likelihood. Prior and likelihood can be expressed easily. In most models they are explicitly stated. The normalization constant, on the other hand needs to be computed as

$$p(x) = \int_{\theta} p(x|\theta)p(\theta)d\theta \tag{1}$$

Above integral can be calculated relatively easily in lower dimension. However, in cases of higher dimension it becomes difficult to compute. This makes posterior distribution even more difficult to calculate leading us to find some approximation techniques to calculate the posterior distribiution. Out of many such techniques, MCMC method turns out to be one of the most efficient techniques to overcome such computational difficulties including the difficulties related to normalization factor.

Sampling Approach: The idea is to draw samples by constructing Markov chain, where the probability of drawing next sample is dependent upon the previous sample drawn. We expect that

the chain will hit, at some point, our desired state or an equilibrium sometimes stationary state which is what we want to sample from. The stationary distribution is our target distribution.

But first off, we need to indetify and define our markov chain. Once we do it, then we simulate a randomly generated sequence and choose some of them as samples that are iid and follow our target distribution. However, for our samples to follow the target distribution we must only consider such states that are far enough from initially generated sequence to have followed, at least in theory, the target distrribution. Therefore, the initially simulated states are discarded from samples. The time period that is needed to reach the stationary state of Markov chain is called burn-in time.

Markov chain, by definition, entails a close sense of dependency and correlation between successive events. So, we keep samples that are relatively far from each other that could be considered as independent. Also, to retain our samples as iid, we also cannot allow to keep all subsequent states after burn-in phase. To achieve this, the gap or lag needed between two states, denoted by 'L' is also considered independent.

The states generated by are seperated by 'L'. Let's consider our Markov chain as

$(X_n)_{n\geq 0} = X_0, X_1, X_2, \ldots$

Burn-in time is denoted by 'T'.

We keep samples

$X_T, X_{T+L}, X_{T+2L}, X_{T+3L}, \ldots\ldots\ldots$

The process can be illustrated clearly by following chart also,

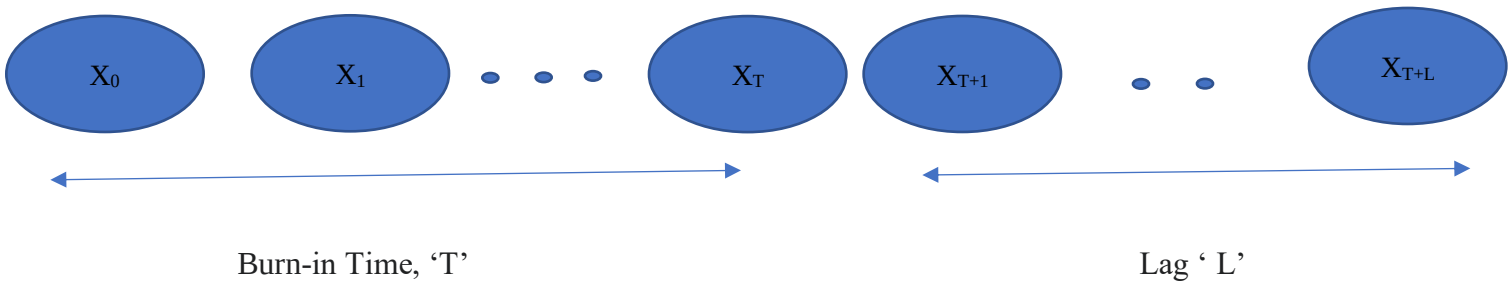


Fig: MCMC sampling showing burn-in time and a lag.

On the left side of the figure, the chain has not hit the stationary state and hence do not follow the target probability distribution. The states that are correlated with X_T will be discarded because we want to generate as independent samples as possible.

The two most famous MCMC algorithms, Gibbs sampler and Metropolis-Hastings whose quick overview I present below. I wrote he following descriptions from various artiles but mostly from the book (Gelman et al., 2013).

1) Metropolis-Hastings Algorithm: It is one of the oldest Markov chain method used by statisticians today. The algorithm is applicable for such probabilistic models from which direct sampling is difficult. It produces a Markov chain whose member's limiting dstribution is our target distribution denoted by $\pi(x)$. It involves picking an arbitrarily

initial value and iteratively accepting and rejecting possible candidate samples drawn from another distribution which is easy to sample.

Let's assume we want to sample from a target distribution P(Θ). However, we know it only upto a normalization constant , let's say, g(Θ).

$$P(\Theta) \alpha\ g(\Theta)$$

We do not know much about normalization constant because it may be practically not integrable. Then the algorith follows following procedures.

i) Selecting an initial value Θ_o.

ii) For large number of iterations, i.e. for i = 1, 2, 3, …….n we repeat followings

a) Draw a candidate Θ* from a proposal distribution, lets say, q(Θ). So, we draw Θ* given the previous iteration Θ_{i-1}.

$$\therefore \quad \Theta^* \sim q\,(\Theta^* | \Theta_{i-1})$$

b) Now we calculate the following ratio, commonly denoted as α.

$$\alpha = \frac{g(\Theta*)/q(\Theta*/|\Theta_{i-1})}{g(\Theta_{i-1})/\,q(\Theta_{i-1}|\Theta*)} = \frac{g(\Theta*)/q(\Theta_{i-1}|\Theta*)}{g(\Theta_{i-1})/\,q(\Theta*|\Theta_{i-1})}$$

c) Check α. If $\alpha \geq 1$ we accept Θ* and set $\Theta_i \rightarrow \Theta^*$.

d) If $0< \alpha <1$ we accept Θ* and set $\Theta_i \rightarrow \Theta^*$ but with some probability α.

In any other case we reject Θ* and set $\Theta_i \rightarrow \Theta_{i-1}$ with probability 1-α.

Since our proposal distribiution q(Θ) is not our target distribiution P(Θ), the steps b and c work as correction. At every step of the chain, we draw a candidate and we decide whether to move ahead or remain in the same position. If the move to the candidate is advantageous i.e $\alpha \geq 1$ then we certainly move. But if the move is not advantageous we stil move in the chain but with probability

α. The candidate generating distribution q(Θ) may or may not depend on the previous value of the Θ. It is obvious that not all candidates that are drawn are accepted and in that case Markov chain stays in the same place as it is possibly for many interations. How often we want to accept or reject the candidate depends on the type of algorithms we use. If we approximate p(Θ) with q(Θ) and draw candidate from that distribution then accepting candidates often is useful. This implies that q(Θ) is estimatimng p(Θ) well. However, we still want q(Θ) to have a larger enough variance than p(Θ) and observe some rejections so that we are assured that q(Θ) is covering our sample space well.

2) Gibb's Sampler: Gibb's sampler, a computer intensive algorithm is another popular statistical tool used widespread today. It is used when the joint distribution is not known explicitly and difficult to sample from. This algorithm is also known as a special case of Metropolis Hastings algorithm. It generates a Markov chain whose value converges towards our target distribution. Unlike in Metropolis Hastings algorithm, proposal candidates are always accepted in Gibbs Sampling.

Let's consider a multivariate probability distribution, P(X, Y) which is basically a function of multiple variables. Because of some mathematical and computational difficulties we do not know how to sample from it directly. However, we have some information about their conditional probabilities, P(X|Y), P(Y|X).

Then Gibbs sampling method becomes really easy to apply in such case when the joint distribution is not known explicitly but the conditional probabilities of each varible is known and easy to

sample from. Like Metropolis Hastings algorithm, it also starts by choosing an initial values for random varible X and Y. Then we sample from the conditional probability P(X|Y).

$P(X|Y) = Y_0$

Next we do sample a new value of Y under condition X_1. We repeat this process upto n-1 interations alternating between new samples from the conditional probability for X and Y under the conditions of given value of random variables.

The steps follows,

1) Initialize X_0 and Y_0
2) For large iterations, for i = 1, 2, 3,……

   sample$X_j \sim p(X|Y_{j-1})$

   and sample $Y_j = p(Y|X_j)$

It is very usual to discard some samples in the beginning during burn-in period. For instance, first 500 samples be ignored and then every 100^{th} samples be averaged casting off the rest. The reason is simply because it takes a while to reach the desired distribution of Markov chain i. e. our stationary distribution. Also, the successive samples are not completely independent. Rather, they have some degree of correlation in the chain. This is why this algorithm can sometimes be used to compute autocorrelation between the samples drawn.

In a nutshell, Gibb's sampling is a MCMC method that draws samples iteratively from the distribution of each variable, conditional on the each variable to estimate a joint distribution

Bayesian Estimation in Infectious Disease Modelling using SIR Model:

Brief description of Covid-19 disease:

We live in an extraordinary time. The COVID-19 disease continues to rage across the globe posing a significant threat to global economy and public health. The disease has shown us that it can simply overwhelm health care system of even well-developed nations (S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, & M. Lipsitch, 2020a). Compounding the problem, new variants of COVID have been detected across the countries. For instance, on 20th September 2021, a new variant of COVID named B117 was 1st found in London UK (Duong, 2021). On the brighter side, a first generation of COVID-19 vaccines are made available to public by several countries thanks to the pharmaceutical giants like Pfizer, Moderna and BioNTech. As of 2021-05-29, a record number of 1.84 billion vaccine doses have been administered worldwide (The New York Times). However, a vast amount of research is being needed to confirm whether the vaccines available now can really stave off these new variants (Mahase, 2021). As the world grapples with new/new variants, it is still important than ever before to predict the dynamics of the pandemic and what factors are contributing to its dynamics in order to come up with effective control measures.

The COVID-19 pandemic is caused by a new strain of virus called severe acute respiratory syndrome coronavirus 2 or (SARS-CoV-2), first identified in December 2019, Wuhan, China (Kissler et al., 2020a). It was declared a pandemic by WHO in March 2020. As of now (30th June

2021) more than 3.95 million people have died worldwide, and more than 182 million cases reported as of (WH0 and Worldometer).

It transmits from an infected person's mouth or nose in small liquid particles called aerosols when they cough, sneeze, speak, sing, or breathe heavily. People get infected with COVID-19 when the virus gets into their mouth, nose, or eyes, which is more likely to happen when people are in direct or close contact (less than 1 metre apart) with an infected individual.

As of now most research and evidence point out that the main way the virus spreads is by respiratory droplets among people who are in close contact with each other.

Studies around the world are underway as to how these aerosols transmit in any specific circumstances, particularly in indoor, crowded and inadequately ventilated spaces, where infected person(s) spend longer periods of time with others, such as restaurants, choir practices, fitness classes, nightclubs, offices and/or places of worship. More studies are being conducted to better understand the conditions in which aerosol transmission is occurring outside of medical facilities where specific medical procedures, called aerosol generating procedures, are conducted.

The other less common way virus spreads after infected people sneeze, cough on, or touch surfaces, or objects, such as tables, doorknobs and handrails. Other people may become infected by touching these contaminated surfaces, then touching their eyes, noses, or mouths without having cleaned their hands first (WHO date 2020, Nov 1).

a) Introduction to Epidemiology Modelling.

The very first noticeable contribution to infectious disease modelling can be traced back to John Graunt(1620-1674), a Londoner and a habersher by profession. *In his book, Natural and Political Observations made upon the Bills of Mortality (1662)*, he tried to present systematically causes of deaths dividing it into different categories such as gender, location, and seasonal variations (Pandey, Chaudhary, Gupta, & Pal, 2020; Snow, 1936). This is regarded as the onset of 'theory of competing risk', an underlying theory in modern epidemiology (Chen, Lu, & Chang).

Another big step came from well-known Swiss physicist and mathematician, Daniel Bernoulli (1700-1782). He demonstrated in his mathematical model that the inoculation against smallpox would have the ability to increase their life expectancy (Hethcote, 2000).

b) From 1920 to present: Birth of compartmental modelling.

The year 1920 marked the beginning of systematic studies into compartmental modelling in epidemiology thanks to the valuable contributions from researchers like William Ogilvy Kermack (1898 – 1970), Anderson Gray McKendrick (1876 –1943), Lowell Reed (1886–1966), Wade Hampton Frost (1880 – 1938). The very famous 'Kermack-McKendrick model' laid the foundations on successfully predicting dynamics of outbreaks (Brauer & Castillo-Chávez, 2001).

Types of mathematical model

a) **Stochastic Model or Non-Deterministic Model**
b) **Deterministic Model**

a) **Stochastic Model**: A quantitative description of a natural phenomenon is called a mathematical model of that phenomenon and a stochastic process is known as a family of random variables X_t, where 't' is a parameter running over a suitable index 'T' (Pinsky & Karlin, 2010). A stochastic model can be explained as a mathematical model that we use to estimate the probability distribution of unobserved random variables by allowing a change in one or more random variables over a specific period. The fact that it is built upon various uncertain factors gives us different answers and estimations. The process is repeated many times under different circumstances to see and assess outcomes. This model is very famously applicable in industries like insurance, stock market and biology.

b) **Deterministic Model**: A deterministic model can be described as a mathematical tool where the input variables change as per specific mathematical equations not as some random variation in variables. In other words, it is something in which all variables can be uniquely determined some parameters of the model and previous state of the variables. Consequently, it depends on the given initial conditions.

When we consider a huge population, in cases a pandemic like Covid-19, this type of model is preferred. A deterministic model is expected to give a satisfactory picture of developing process such as an epidemic or growth of the population so long as the population is sufficiently large (Rushton & Mautner, 1955). In deterministic model, every member of the population is assigned to a different compartment representing a specific stage of the disease. Then the transition from one compartment to another compartment is represented by differential equations provided that the population is differential over time and the whole process is deterministic.

Examples of Deterministic Models: One of the simplest deterministic models in epidemiology is the SIR model. By adding and removing more compartments into SIR, one can develop its derivatives like SIS, SEIR, SEI, and several others. To achieve more accuracy of the dynamics of any infectious disease we can tweak and add more compartments.

Introduction to SIR Model: The simplest compartmental model that had the ability to effectively predict the pandemic like covid-19 was first presented by researchers Willian Ogilvy Kermack(1898-1970) and Anderson Gray McKendrick(1876-1943) over the years of 1927, 1932 and 1933 in there different articles (Kermack & McKendrick, 1927). Their seminal work laid the foundations for later developments of mathematical models of various infectious diseases. The SIR model, however, ignores the heterogeneity of the susceptible population distribution, gives us an overview of the dynamics of the disease. Later Kendall 1957 (Kendall, 2020) model took spatial component into account to address the non-local transmission which I will briefly discuss later.

The SIR model (without vital dynamics): In 1927, researchers, Kermack and McKendrick first came up with a model in which population was compartmentalized into three separate compartments, S(t) for Susceptible, I(t) infected for R(t) Removed. In this model, it is assumed that all population is homogenous, and members are equally susceptible to the disease. Therefore,

S(t) →denotes the members who have not been infected with the disease at a time 't'.

I(t) →denotes the members who have been infected with the disease and are prone to

R(t) →denotes the members of the population who have already been infected and removed from our compartment either due to getting immunity or due to death.

Figure: 1 (Compartmental Flow in SIR model)

In this model we assume the followings:

i)The total population of three compartments let's say, 'N' remains the constant over time 't'.

ii) That the total time period of the disease is same as the total time period of infection with constant transmission and recovery rate. And constant rate of increase in infection is proportional to the contact between infected and susceptible within the population.

iii) Since we are considering SIR model without vital dynamics, we do not consider natural birth and death within the population.

Mathematically we can write the above assumptions as following ordinary differential equations (ODE).

The rate of change of susceptible with respect to time can be given as,

$$\frac{ds}{dt} = -\alpha IS \tag{1}$$

Where α=constant rate of contact between number of infected and number of susceptible individuals.

Negative sign in the equation indicates that number of susceptible are decreasing as people become infected with the disease.

Similarly, the rate of change of infected individual can be given as,

$$\frac{dI}{dt} = \alpha IS - \beta I \tag{2}$$

The first term in the equation is same and bears positive sign because now the number infected individuals are increasing.

The second term represents that infected individuals recover or die from the disease at constant rate ‘β’.

$$\frac{dR}{dt} = \beta I \tag{3}$$

This term is the same as the last term of equation (2). It represents that the individuals that are lost from the infected compartment are now moving to ‘Removed’ compartment.

Let’s assume initially at the beginning of epidemic,

S=So because the whole population is susceptible to the disease, especially in case of never seen pandemic like Covid-19.

I=Io, R=0

From equation (i), we can see ‘S’ is always decreasing.

So, S≤So

Now we put this value of ‘S’ in second equation to see rate of infection.

$$\frac{dI}{dt} \leq I(\alpha So - \beta) \tag{4}$$

Now, for the disease or epidemic to occur, it ultimately boils down to the sign of the second term in (4). It has to be positive,

Therefore, $\alpha S_o - \beta > 0$

Or, $S_o > \frac{\beta}{\alpha}$

Or, $S_o > \frac{\beta}{\alpha} = \frac{1}{p}$ (assume) (5)

From (v), I introduce a new term, denoted as 'Ro'.

$$Ro = \frac{\alpha So}{\beta} > 1 \quad (6)$$

This new term, 'Ro' is called basic reproduction/replication number or ratio.

This term has a great significance in our model and in epidemiology at large. It basically tells us how fast a disease can spread. Precisely, it represents average number of secondary infected people caused by single initial individual during his/her infection period. It depends upon duration of infection, the probability of infecting a susceptible person during one contact, and a number of susceptible people contracted per unit time (Dietz, 1993). For this reason, it may vary not only for different kind of infectious diseases but for same disease in a different kind of population.

Also, the time between contacts is $T_c = \alpha^{-1}$ and time until an individual is removed from the compartment is $T_r = \beta^{-1}$. From these values we can also come up with average number of contacts made by an infected individual before he/she is removed as T_r/T_c.

If we divide equation (1) by equation (3), separate the variables and integrate, we get followings.

$$S(t) = S(0)\, e^{-Ro(R(t)- R(0))/N} \quad (7)$$

Where S(0) and R(0) are initial value of susceptible and removed populations. If we write

$s_0 = S(0)/N$ for initial condition and $s_\infty = S(\infty)/N$ and $r_\infty = R(\infty)/N$ for susceptible and removed population the as time 't' $\rightarrow\infty$ we get

$$s_\infty = 1- r_\infty = S_0\, e^{-R0(r_\infty - r_0)}$$

The infectious compartment empties at the end 't'$\rightarrow\infty$. This equation has a solution in terms of Lambert W function (Teles, 2020) as $s_\infty = 1- r_\infty = - R_0^{-1}\, W(- S_0\, R_0\, e^{-R_0\,(1-r_0)})$ (8)

We will have following conditions to describe how the disease is moving across population depending on the value or R_0.

Case I: Ro < 1

Primary infected person infects fewer than one person and eventually the disease will die out.

Case II: Ro = 1

Primary infected person will infect exactly one secondary individual. In such situation the disease is called endemic. It will spread throughout the population, but will not increase or decrease, however.

Case III: Ro > 1

Primary infected individual will infect, on average, more than one secondary individual and disease can spread very quickly. The situation is called 'epidemic'. In case of seasonal flu, it is estimated that the median R_0 to be somewhere around 1.28 (Biggerstaff, Cauchemez, Reed, Gambhir, & Finelli, 2014). In case of Covid-19, the mean value of 'Ro' is estimated to be 3.38 ± 1.40 with a range of 1.90 to 6.49 (Alimohamadi, Taghdir, & Sepandi, 2020). Also, the pooled R_0 value for COVID-19 was estimated to be 3.32 (95% confidence interval, 2.81 to 3.82) (Alimohamadi et al., 2020).

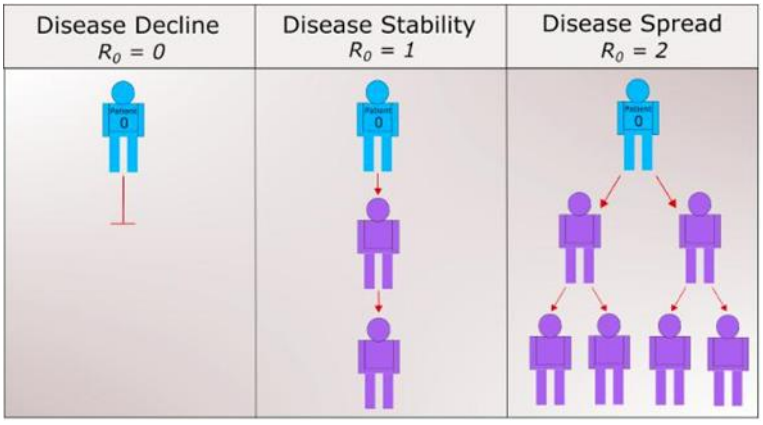


Fig: Depiction of spread of disease as per R_0 value

Source: (Pandey et al., 2020)

Effective Reproduction Number (R_E): In a real case scenario, a whole population will never be completely susceptible to an infection. For instance, some portion of population will be immune to disease because of the immunity granted by prior infection. For this reason, not everyone will get infected in practical world and average number of secondary infections per infection will be less than the basic reproduction number. Therefore, effective reproduction number is defined as the average number of secondary infection cases in a population made up of both susceptible and non-susceptible people.

Mathematically, it can be estimated as the product of basic reproduction number R_0 and portion of susceptible population.

$$R_E = R_0 . S(t) \quad (7)$$

Herd Immunity: Herd immunity is a stage during a period in pandemic when a significant number of people become immune either by vaccination or by self-immunization after contracting the disease, making for an overall protection of remaining susceptible population. The greater the number of immune people, the lower the probability that remaining people will meet infected individual and contract the disease. This, in turn makes even more difficult for disease to spread if sufficiently large number of people are immune as cycle of infection is broken.

The heard immunity threshold can be defined as the portion of population needed to be immune to make disease stable. Or, the point at which the proportion of susceptible population falls below a minimum needed for transmission is known as herd immunity threshold (Anderson & May, 1985). After herd immunity stage, each new case leads to, at most, a new single case ($R_E = 1$) and hence making the infection stable.

∴ Heard Immunity Threshold (HIT) = $\frac{R_0-1}{R_0}$

After HIT is surpassed, R_E becomes less than 1. i.e R_E <1 and the number of cases of infection decreases. This constant gives us an effective measure for infectious disease expert to evaluate and implement eradication policies. Figure below depicts HIT in SIR model.

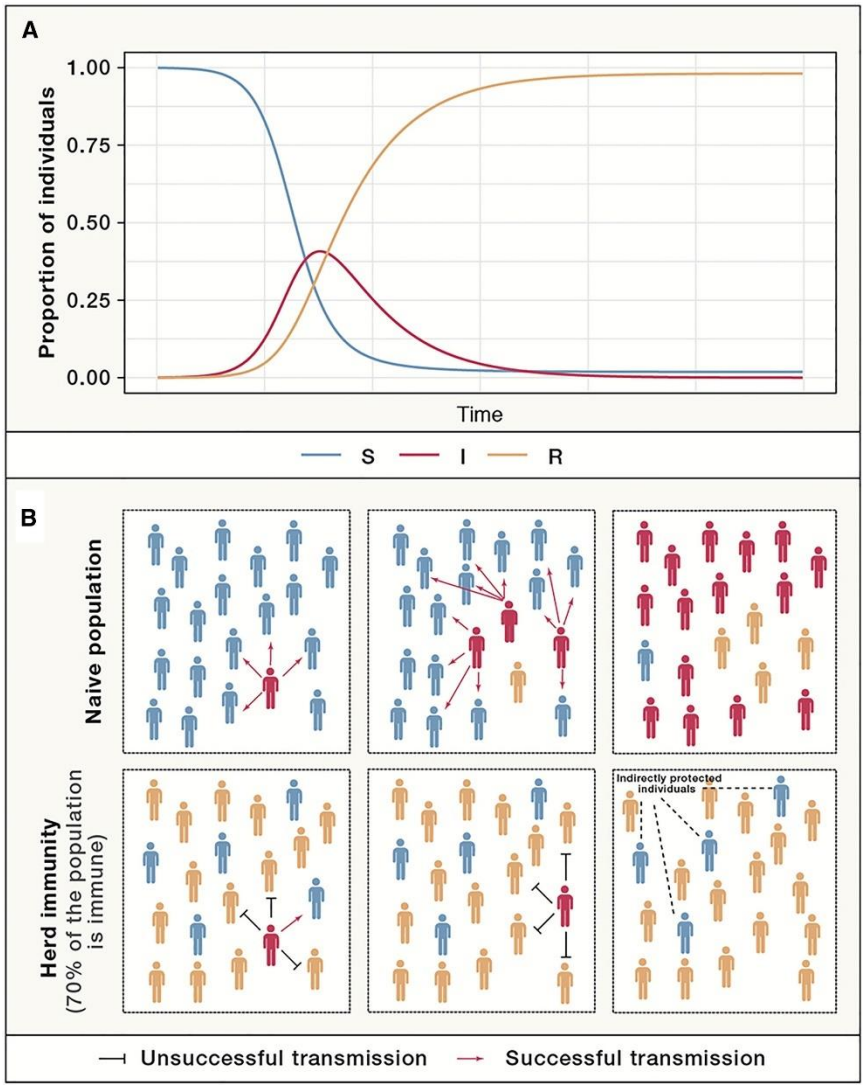
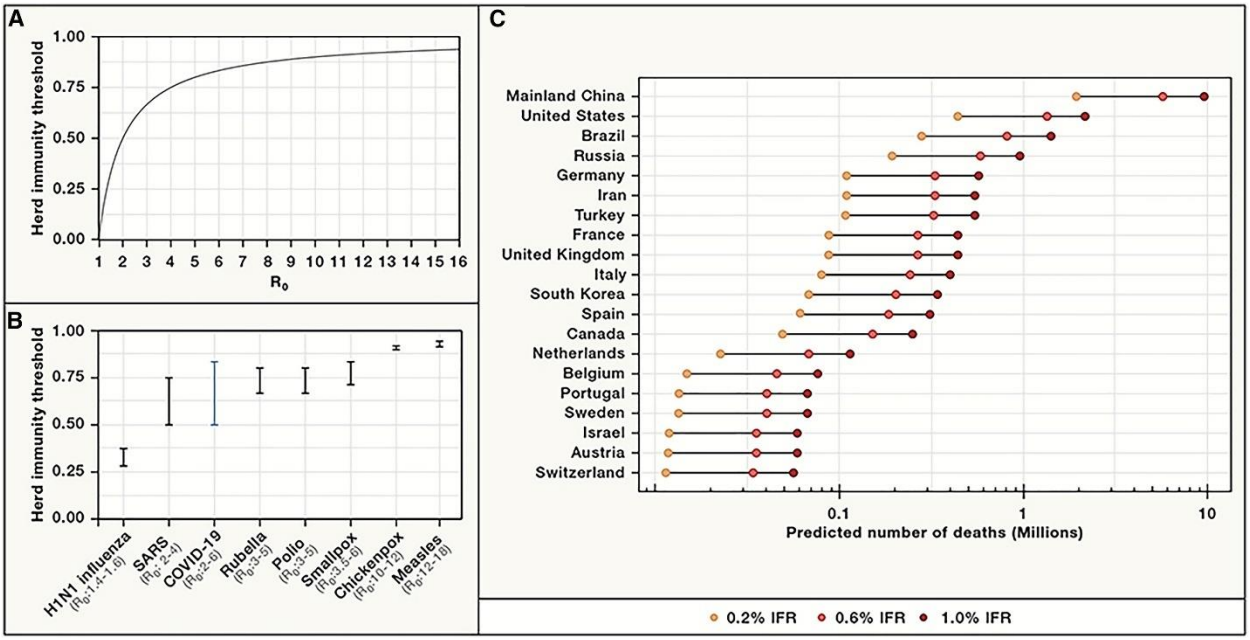


Fig: 2 Herd Immunity using SIR model for totally immunizing infection with an $R_0 = 4$. Source: (Randolph & Barreiro, 2020)

In above figure A, after an introduction of a new case the portion of infected individuals denoted by red line increases rapidly until it reaches its peak point corresponding to herd immunity threshold (HIT). This point represents the stage where newly infected individual will infect fewer than 1 individual as sufficiently large number of people have now become immune averting further infection cycle.

And figure 2B depicts the disease propagation dynamics. In naïve population above, an outbreak can quickly occur but in the case of herd immunity, the virus does not spread and prevail in the population thereby stopping the further infection.



Source: (Randolph & Barreiro, 2020)

Fig: 3 The health burden of COVID-19 if herd immunity is achieved without vaccination.

Above chart 3A represents the herd immunity threshold i.e. the portion immune population needed to reach herd immunity stage (y-axis). As R_0 increases, the population proportion needed to reach herd immunity also increases. Figure 3B depicts the basic reproduction ratio for different infectious diseases and in fig 3C we see expected number of deaths for 20 countries under uniform herd immunity threshold of 67% i. e. (R_0 = 3) and overall COVID infection fertility rates, 'IFR' of 0.2%, 0.6% and 1.0% are assumed.

Calculating the Maximum Number Infected People at any given time:

Let's divide equation (i) by (ii) and simplifying.

We get, $\frac{dI}{dS} = \frac{\alpha IS - \beta I}{-\alpha IS} = -1 + \frac{\beta}{\alpha S} = -1 + \frac{1}{pS}$ (Assume) (8)

In equation (vii), $p = \frac{\alpha}{\beta}$, is defined as contact ratio.

Further simplifying (8)

Or, $dI = dS\left(-1 + \frac{1}{pS}\right)$ (9)

Integrating above equation

We get,

$$I + S = \frac{1}{p} lnS$$

But we have our initial conditions as

S= So (initially)

I= Io (initially)

R= 0 (initially)

And $\frac{d}{dt}(S + I + R) = 0$

$$S + I + R = So + Io \quad (10)$$

So, applying those initial conditions we get our equation as

$$I + S - \frac{1}{p} lnS = Io + So - \frac{1}{p} lnSo \quad (11)$$

Now from our equation (8), it can be seen that '$\frac{dS}{dI}$' will be minimum when S= $\frac{1}{p}$

Putting this value in our eq^{n} (9) and solving we get I_{max} as

$$I_{max} = Io + So - \frac{1}{p}\{(1 + ln(pS_o)\} \quad (12)$$

So, from (11) our I_{max} is equivalent to whole population some minus some number given by the term $\frac{1}{p}(1 + ln(pS_o)$

If we rename this term as f(p) and try to plot it against 'p'. Then the graph may look like following.

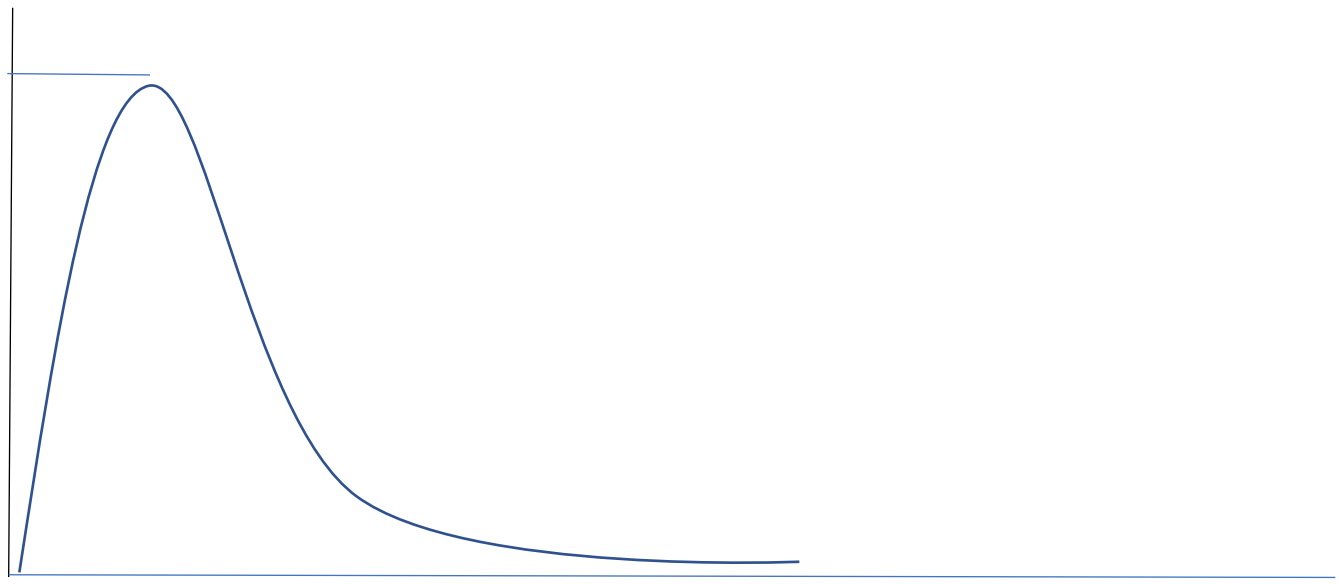


Fig: 4

From this graph, it is obvious that for high value of ‘p’, our function f(p) has smaller value. Using this result, we can say, from (12), that I_{max} will also be high depending upon the value of contact ratio. For very high value of ‘p’ almost all the population will catch the disease. Therefore, the question of maximum number people that can catch the disease depends on only one quantity, the contact ratio. So, for a pandemic like Covid-19, the contact ratio is relatively high and hence huge number is people are getting infected. All the measures that we are told to follow by health officials such as, social distancing, frequent hand washing, putting on a mask and so forth are directed towards lowering this term contact ratio so that less and less people get infected.

Calculating total number of people that may catch the disease: For this, we need to know what it means for disease to end.

The disease will end when number of infected people go to zero and, hence ending the outbreak.

At the end of the outbreak, our ‘R’ compartment will give us how many people recovered or died from the disease.

To calculate this, let’s go back to eq^{n}(10) and rewrite it as following

$$R_f = S_f + I_o + S_o$$

Where, $R_f \rightarrow$ total number of people recovered or died from the disease.

$S_f \rightarrow$ total number of susceptible remained at the end of the disease.

Taking help from eq^n (11) and rearranging it for this case, I=0.

We get,

$$S_f - \frac{1}{p} ln S_f = Io + So - \frac{1}{p} ln So$$

Like before if we try to plot S_f against p, this time our graph may look like following.

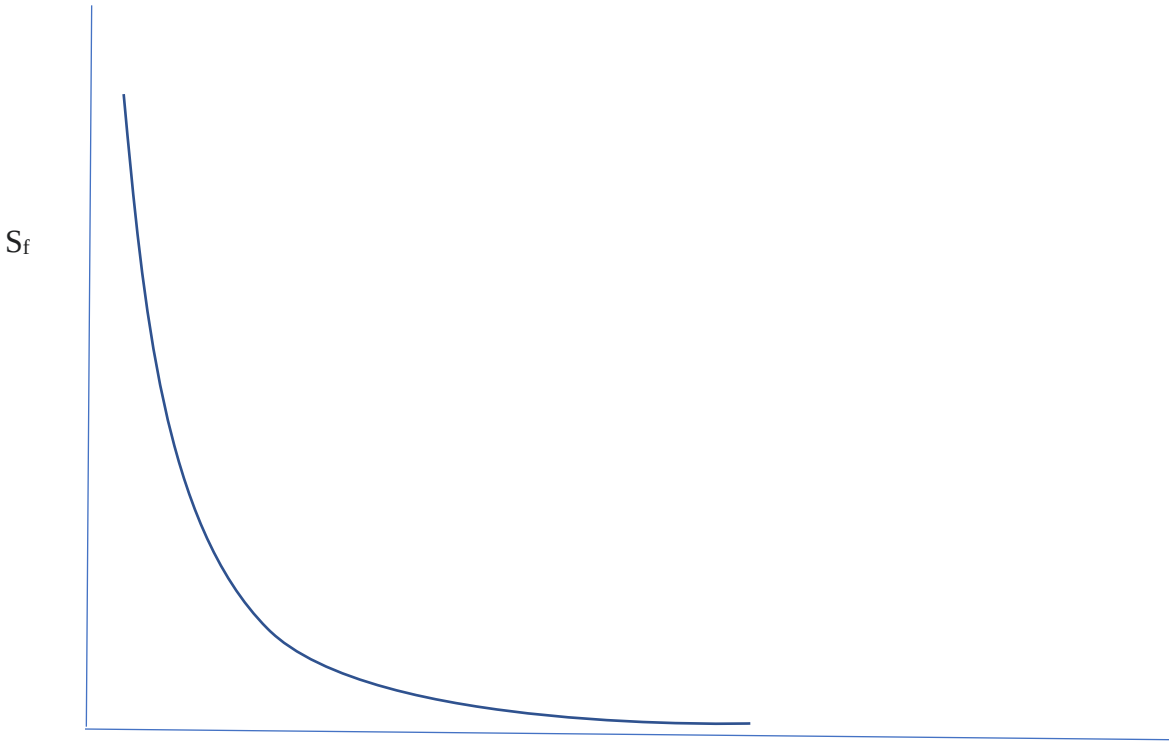


Fig: 5

From graph it seems obvious that for large value of q, S_f is small again. This concludes us again that almost all population have the chance of getting infected should the value of 'q' is very high.

Conclusion:

For sufficiently large value of 'p' and hence the basic reproduction number 'R_o' the disease will spread very rapidly, and a pandemic may occur.

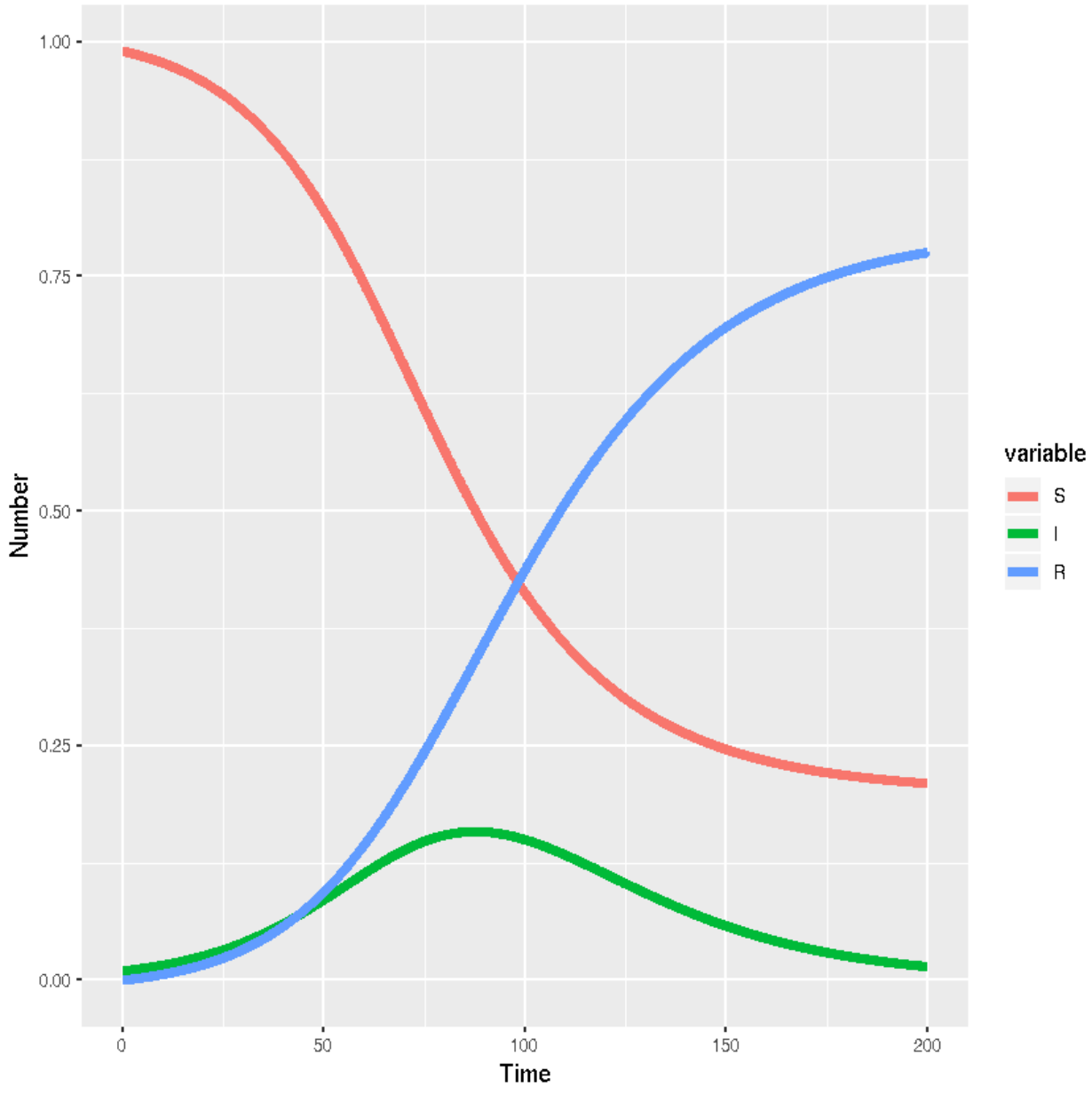


Fig: 6 ggplot2 of SIR model using deSolve R package.

Source: Simon Frost (Date: 2018-07-12)

This figure also depicts the disease dynamics across different population compartments. As the time passes, the susceptible population portion falls rapidly (red line) and at the same time, infected population increases until it reaches its peak (green line). The infected individuals denoted by blue line also increases as more and more people are removed either recovery or by death.

Pandemic Wave and Seasonality Factor:

As yet there is no hard and fast definition of what a pandemic wave is. However, a disease wave simply implies increasing number individuals who get infected which also can be shown in peaks and declines in graphs. Even after a significant decline or after a stagnation, severe outbreaks are possible. By carefully analysing the historical outbreaks of different infectious diseases from the past, we can predict how COVID-19 will behave over the years. Spread of such respiratory viral infectious diseases like COVID-19 have many dependencies. One of the important such factors I am going to mention is seasonality or seasonal amplitude. Since the inception of this outbreak, it was widely believed that COVID-19 will, like other respiratory viral infection, show some seasonality dependency (Liu et al., 2021). Several seasonal coronaviruses such as 229E or HKU1 showed peak infection from around December to March (Monto et al., 2020) in the US and elsewhere (Chakhunashvili et al., 2018).

Researches have further shown that transmission of COVID can be dependent on several seasonality factors like temperature and humidity (Altamimi, Ahmed, & Health, 2020) that determines the survival of the virus and transmission routes. (Sajadi et al., 2020) found there is significant outbreaks of COVID-19 are found in the regions with mean temp between 5 degree C and 11 degree C, combined with low humidity (3- 6g/kg).

Pandemic waves and seasonal dynamics are also dependent upon other factors such as human immunity across different regions. As more and more people get infected and develop immunity, the speed of the wave with which it spreads also slows down and eventually stops as it can no

longer find susceptible individual. Studies tell us that it will eventually become other seasonal coronaviruses (S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, & M. J. S. Lipsitch, 2020b). However, (Britton, Ball, & Trapman, 2020) found that 43% to 60% of the US population would be needing to get immune to reach herd immunity stage.

Travelling waves for SIR Model:

Here our aim is to find out how fast the disease is spreading among populations. The reason we call wave solutions is that the disease is spreading like a wave with some constant speed infecting our susceptible population.

In SIR model mentioned above, we allowed to vary three population categories with time. Here, we want to vary population with space also as people start move around the compartment and thereby introducing the spatial dependence of our three-population category. Also, we assume that the susceptible and removed population do not move. Only infected population move. This gives us a more realistic approach to build the model. For this, we again start with same three equations as we have it before.

$$\frac{ds}{dt} = -\alpha IS \qquad \text{(i)}$$

$$\frac{dI}{dt} = \alpha IS - \beta I \quad \text{(ii)}$$

$$\frac{dR}{dt} = \beta I \quad \text{(iii)}$$

But our population is now dependent on both time and space so we will use partial differentiation. Our equations now become,

$$\frac{\partial S}{\partial t} = -\alpha IS \quad \text{(iv)}$$

$$\frac{\partial I}{\partial t} = \alpha IS - \beta I \quad \text{(v)}$$

$$\frac{\partial R}{\partial t} = \beta I \quad \text{(vi)}$$

As mentioned above, we have assumed that only the infected individuals move around the compartment with some constant diffusion rate 'D' spreading the disease leaving 'S' and 'R' category at rest. So, equation (iv) is modified into following.

$$\frac{\partial I}{\partial t} = \alpha IS - \beta I + D\frac{\partial^2 I}{\partial x^2} \quad \text{(vii)}$$

Let's start by introducing a new variable 'z' as follows,

$$z = x - vt$$

Putting this new variable and solving, we end up with following,

$$\frac{d^2S}{dZ^2} + v\frac{dI}{dZ} + I(S - \frac{1}{Ro}) = 0$$

$$v\frac{dS}{dZ} - IS = 0$$

Using boundary values and nondimensionalization technique and solving above equations we get that for travelling wave solutions to exist we must have,

$$V \geq 2\sqrt{(1-1/R_0)} \qquad (*)$$

And the minimum value is,

$$V = 2\sqrt{(1-1/R_0)} \qquad (**)$$

Above equation gives the travelling wave speed of the disease with which it is propagating through the population. We already know that for epidemic to occur value of R_0 must be greater than 1 which also justifies our above wave equation.

The SEIR Model:

The differential equations that govern our new SEIR model will have a new compartment called, 'exposed' denoted by 'E'. This compartment entails the population proportion that has been exposed to infected individuals and potentially may be more prone to the disease. Like any other vector borne disease, Covid-19 has its own incubation period or period of latency. The median incubation period for Covid-19 as shown by early research (Lauer et al., 2020) is found to be 5 days with mean incubation period 5.2 days. Those findings were also in line with later research. However, for other vector borne diseases like chicken pox the incubation period is found to be more than 11-12 days. In the case of Covid-19, the latent phase as stated above, after a person acquires the disease but still now infectious is added to our previous SIR model as 'E' i.e. the exposed population. The flow chart below will tell us clearly about the dynamics of SEIR model without vital dynamics.

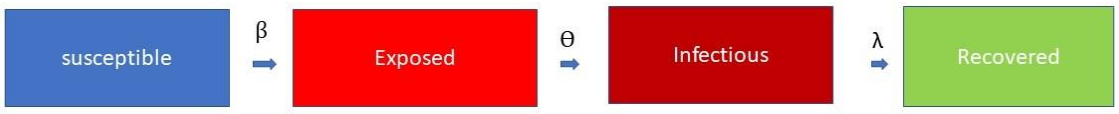


Fig: Compartmental flow in SEIR Model

The differential equations are given as follows,

$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \Theta N$$

$$\frac{dI}{dt} = \Theta E - \lambda I$$

$$\frac{dR}{dt} = -\lambda I$$

Where, N = S + E + I + R is the total population.

In this model, the incubation period of period of latency delays initial infections so we observe slower initial growth in secondary infections. Later, we see faster growth than that of SIR model which has no latency or incubation period. Since we are not considering vital dynamics our basic reproduction number R_o remains same.

$$Ro = \frac{\beta}{\lambda} \qquad \text{(vii)}$$

After slow initial outbreak, the susceptible population is depleted by the disease until there are not enough individual to spread the disease. The shorter the incubation period, the quicker the susceptible population is depleted by epidemic. Usual SEIR Model without any social interventions looks like graph shown below.

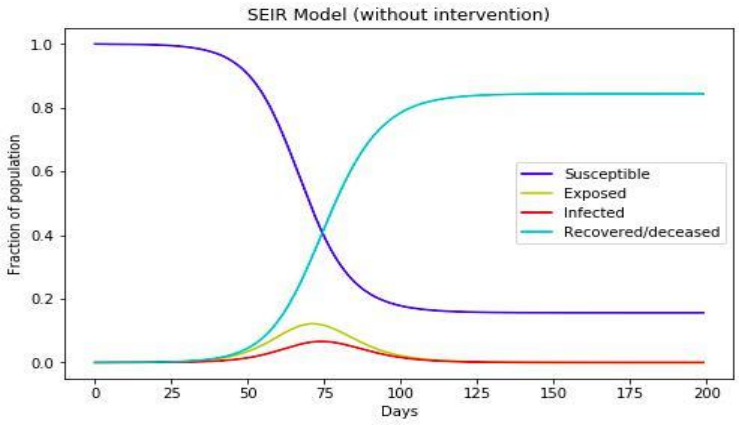


Fig: SEIR model using $R_0 = 4$ and S to be 70% of India's population

Source: (Pandey et al., 2020)

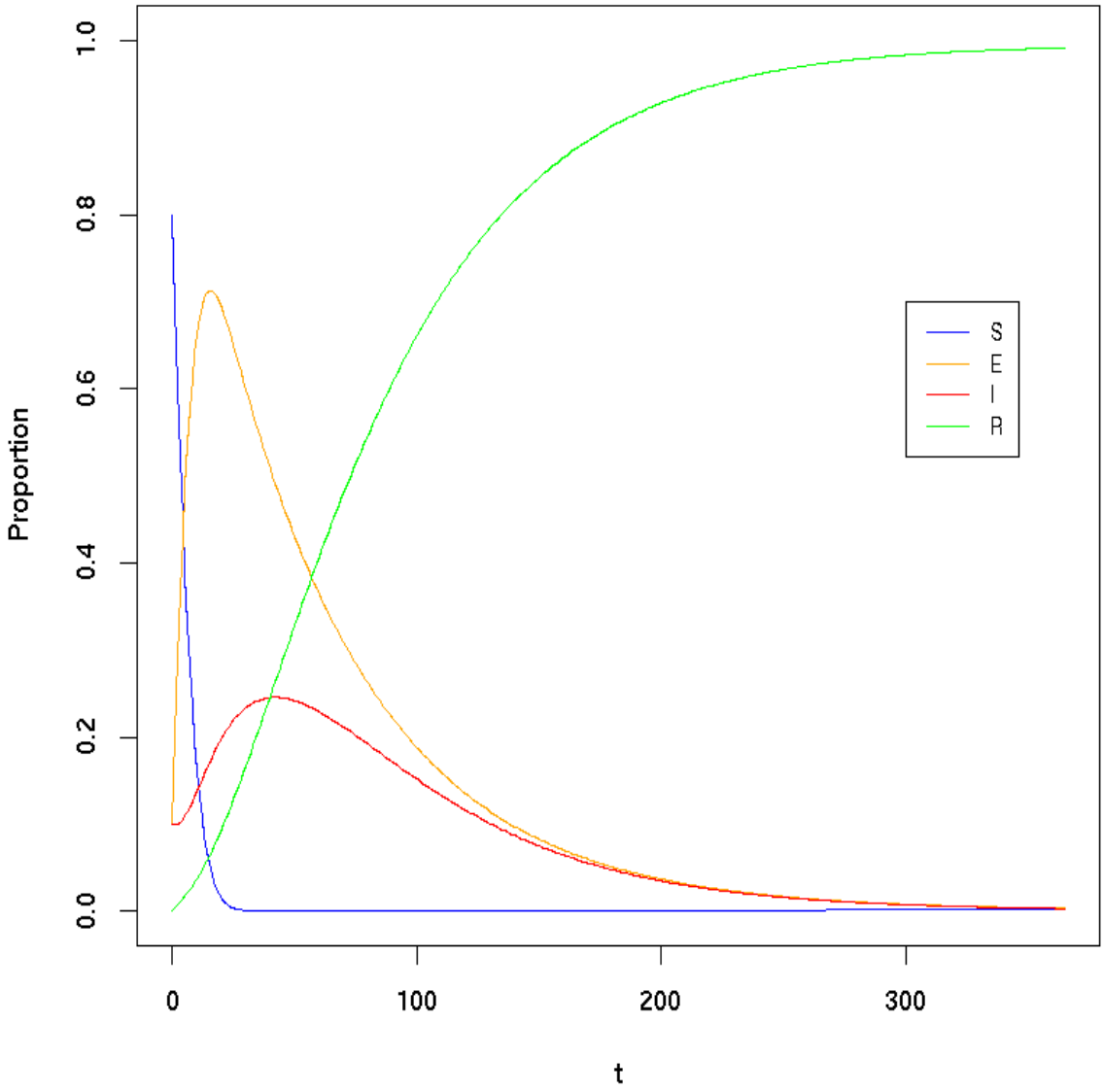


Fig: 7

SEIR Model dynamics using

desolve R package.

Source: Lloyd Chapman (Date: 2018-10-01)

(http://epirecip.es/epicookbook/chapters/sir/r_desolve)

Estimating Model Parameters using SIR model:

Here our aim is to show how the model parameters (SIR) estimation techniques in compartmental modelling work. In other words, we would like to show, in the case of infectious disease, how we use MCMC algorithm, to estimate the so called 'basic reproduction ratio-R' and other valuable transmission coefficients. For this we initially wanted to use real COVID-19 data from various cities within Norway. However, the MCMC algorithm for SIR model did not work perfectly with the data as we observed glitch in our inference table as clearly evidenced by much variation in the value of $\hat{R}$ implying our Markov chains are not in sync with one another. The reason maybe MCMC algorithm could not sample from huge population size, on the scale of (5.3m) and our sample being too small and our model being too simplistic. At the end, we decided to show the same estimation techniques using stan codes from (https://mc-stan.org/users/documentation/case-studies/boarding_school_case_study.html). For this purpose, we followed following steps

i) First, we fit the model in 'R'.

ii) check the inference

iii) check the model.

For this propose we relied on data of influenza A(H1N1) in 1978 at a British Boarding school available on R package outbreaks. It provides us the daily number of hospitalized students. The infection lasted for 14 days. There were in total 763 male students out of which 512 became sick

with infection. The infection was started by one male individual and it lasted from 22 January to 4th February.

We first begin by installing and loading library 'outbreaks' in R. The following R codes are developed by (Grinsztajn et al., 2020) and taken from (https://mc-stan.org/users/documentation/case-studies/boarding_school_case_study.html) page. The Stan codes uses MCMC algorithms to check whether our inference is reliable. The MCMC being one of the finest algorithms tell us by using simulations techniques how can we make sure our model, our priori distribution and inference are correct and are in accordance with observed data. Not only that, Stan, having MCMC as crux of Bayesian estimation technique, provides a way to check the convergence and accuracy of Markov chains so that the inference we get is trustworthy enough.

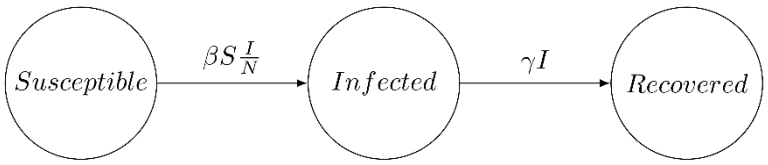


Fig: SIR Model, source (Grinsztajn et al., 2020)

Where,

S(t) = no. of susceptible

I(t) = no. of infected

R(t) = no. of recovered or dead

β = constant rate of contact between of infectious people

γ = constant recovery rate.

Our goal is to estimate these parameters and draw inference using MCMC techniques.

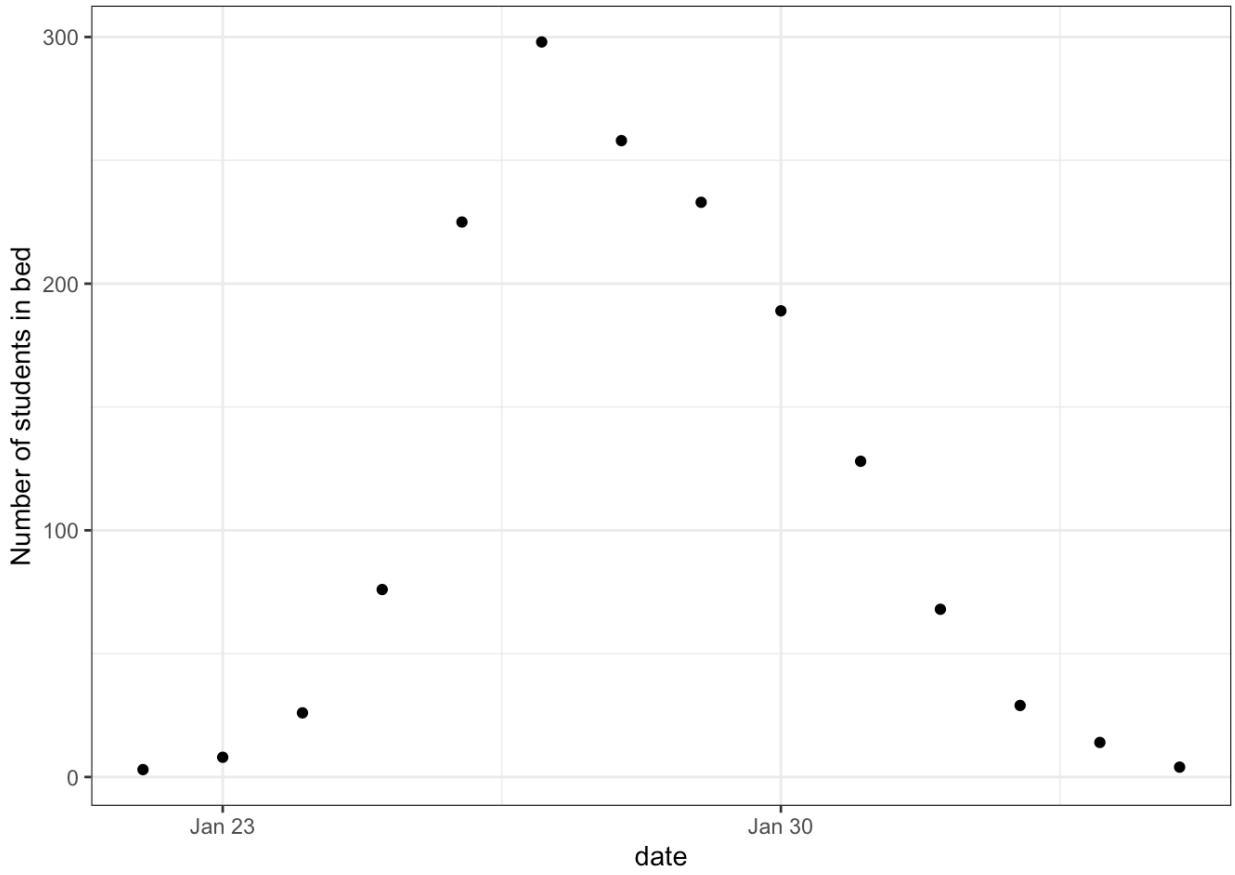
We begin by,

```
library(outbreaks)

head(influenza_england_1978_school)
```

```
##         date in_bed convalescent
## 1 1978-01-22      3           0
## 2 1978-01-23      8           0
## 3 1978-01-24     26           0
## 4 1978-01-25     76           0
## 5 1978-01-26    225           9
## 6 1978-01-27    298          17
```

```
theme_set(theme_bw())
ggplot(data = influenza_england_1978_school) +
  geom_point(mapping = aes(x = date, y = in_bed)) +
  labs(y = "Number of students in bed")
```

If we run above codes, we get the following graph. This graph basically shows us the number of students in hospital bed over the span of Jan 22 to Feb 4.



1st Step: Fitting the SIR model in R:

```
# time series of cases
cases <- influenza_england_1978_school$in_bed  # Number of students in bed
library(outbreaks)
library(rstan)
# total count
N <- 763;

# times
n_days <- length(cases)
```

```
t <- seq(0, n_days, by = 1)
t0 = 0
t <- t[-1]

#initial conditions
i0 <- 1
s0 <- N - i0
r0 <- 0
y0 = c(S = s0, I = i0, R = r0)

# data for Stan
data_sir <- list(n_days = n_days, y0 = y0, t0 = t0, ts = t, N = N, cases = cases)
# number of MCMC steps
niter <- 2000
```

Now we compile the model saved as ‘sir_negbin.stan’ as following,

```
model <- stan_model("sir_negbin.stan")
```

The ‘sir_negbin.stan’ file contains following codes,

```
functions {
 real[] sir(real t, real[] y, real[] theta,
            real[] x_r, int[] x_i) {

    real S = y[1];
    real I = y[2];
    real R = y[3];
    real N = x_i[1];
```

```
    real beta = theta[1];
    real gamma = theta[2];

    real dS_dt = -beta * I * S / N;
    real dI_dt =  beta * I * S / N - gamma * I;
    real dR_dt =  gamma * I;

    return {dS_dt, dI_dt, dR_dt};
  }
}
data {
  int<lower=1> n_days;
  real y0[3];
  real t0;
  real ts[n_days];
  int N;
  int cases[n_days];
}
transformed data {
  real x_r[0];
  int x_i[1] = { N };
}
parameters {
  real<lower=0> gamma;
  real<lower=0> beta;
  real<lower=0> phi_inv;
}
transformed parameters{
  real y[n_days, 3];
```

```
  real phi = 1. / phi_inv;
  {
    real theta[2];
    theta[1] = beta;
    theta[2] = gamma;

    y = integrate_ode_rk45(sir, y0, t0, ts, theta, x_r, x_i);
  }
}
model {
  //priors
  beta ~ normal(2, 1);
  gamma ~ normal(0.4, 0.5);
  phi_inv ~ exponential(5);

  //sampling distribution
  //col(matrix x, int n) - The n-th column of matrix x. Here the number of infected people
  cases ~ neg_binomial_2(col(to_matrix(y), 2), phi);
}
generated quantities {
  real R0 = beta / gamma;
  real recovery_time = 1 / gamma;
  real pred_cases[n_days];
  pred_cases = neg_binomial_2_rng(col(to_matrix(y), 2), phi);
}
```

The next step is to run MCMC. Here we try to run at least 4 Markov chains.

```
fit_sir_negbin <- sampling(model,
                data = data_sir,
                iter = niter,
                chains = 4,
                seed = 0)
```

2nd STEP: Checking the Inference:

Here, we aim to estimate our model parameters, beta, gamma, basic reproduction ratio 'R_0' and recovery time. For this we do following coding.

```
pars=c('beta', 'gamma', "R0", "recovery_time")
```

```
print(fit_sir_negbin, pars = pars)
```

When we run these codes, we get the following results in R.

```
## Inference for Stan model: sir_negbin.
```

```
## 4 chains, each with iter=2000; warmup=1000; thin=1;
## post-warmup draws per chain=1000, total post-warmup draws=4000.
##
##              mean se_mean   sd 2.5%  25%  50%  75% 97.5% n_eff Rhat
## beta         1.73    0.00 0.05 1.63 1.70 1.73 1.77  1.84  2883    1
## gamma        0.54    0.00 0.04 0.46 0.51 0.54 0.57  0.63  2874    1
## R0           3.23    0.01 0.27 2.75 3.05 3.21 3.38  3.84  2812    1
## recovery_time 1.86    0.00 0.16 1.58 1.76 1.86 1.95  2.19  2828    1
##
## Samples were drawn using NUTS(diag_e) at Mon Feb  1 01:02:17 2021.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Above results provide us much needed information as to the inference we draw is reliable or not. The quantiles give us data to check our inference. The beta 'β' was defined as the average number of contacts per person per unit time is found to be 1.73. This implies that, on an average a person has contacted more than single individual at a time leading a way for epidemic to occur.

Likewise, another transmission constant gamma 'γ' is found to be 0.54. The most overriding discovery- basic reproduction ratio, 'R_0' is found to be significantly high – 3.23. This implies that, on average, one single individual transmitted the disease to more than three persons at a time during the span of epidemic. This number tells the severity of the outbreak in epidemiology and has great significance. Further, the average recovery time from the disease is found to be 1.86 days.

On the far-right side of above table Rhat, '$\hat{R}$' is all same – 1. This means that our 4 Markov chains agree with one another. If there was a significant variation in the values of $\hat{R}$, it would mean our

Markov chains are not synchronized with one another. Also, the low value of effective sample size 'n_{eff}' and high value of '$\hat{R}$' would indicate poor mixing of chains.

To plot marginal posterior densities and reaffirm our Markov chains are in perfect sync with each other we run following codes,

```
stan_dens(fit_sir_negbin, pars = pars, separate_chains = TRUE)
```

Wh

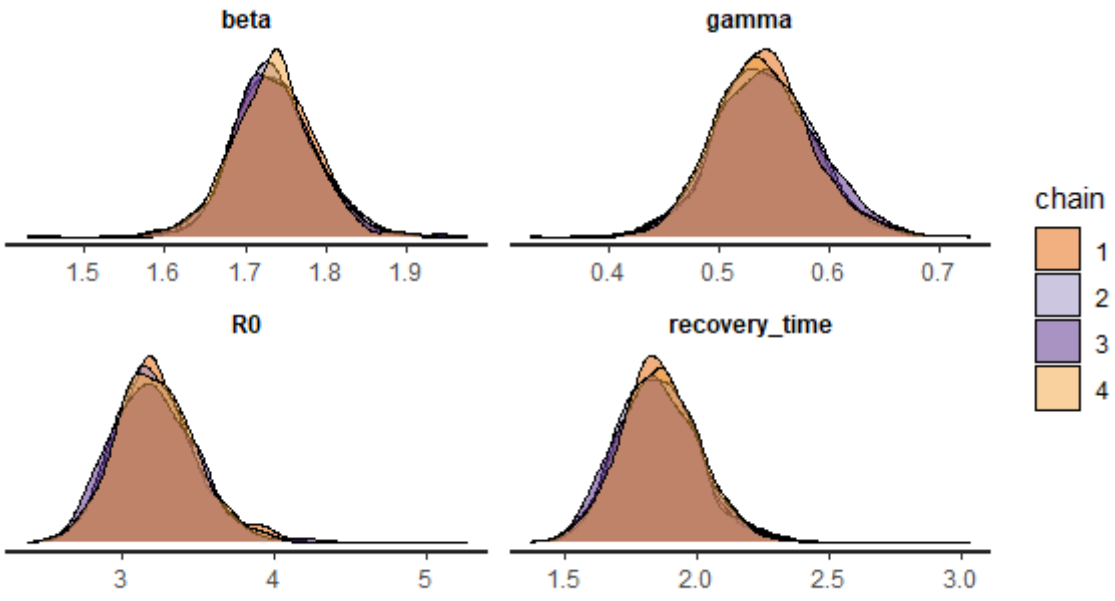


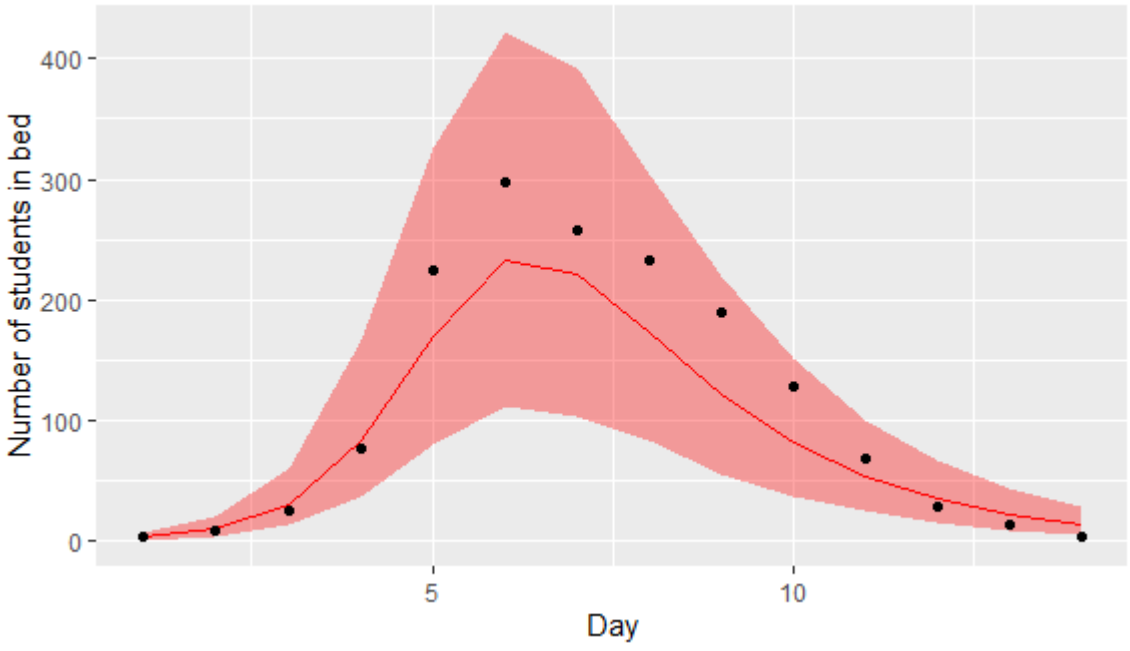
Fig: Marginal Posterior densities of model parameters

3rd Step: Check the Model.

Further, we check our model whether it is best fitted observed data. The method is also called posterior predictive checks. We make sure whether model is giving a befitting answer to the data that also captures the variation within data. For this we do following coding,

```
#Define color
# define color
c_posterior <- "red"
smr_pred <- cbind(as.data.frame(summary(
  fit_sir_negbin, pars = "pred_cases", probs = c(0.05, 0.5, 0.95))$summary), t, cases)
colnames(smr_pred) <- make.names(colnames(smr_pred)) # to remove % in the col names

ggplot(smr_pred, mapping = aes(x = t)) +
  geom_ribbon(aes(ymin = X5., ymax = X95.), fill = c_posterior, alpha = 0.35) +
  geom_line(mapping = aes(x = t, y = X50.), color = c_posterior) +
  geom_point(mapping = aes(y = cases)) +
  labs(x = "Day", y = "Number of students in bed")
```

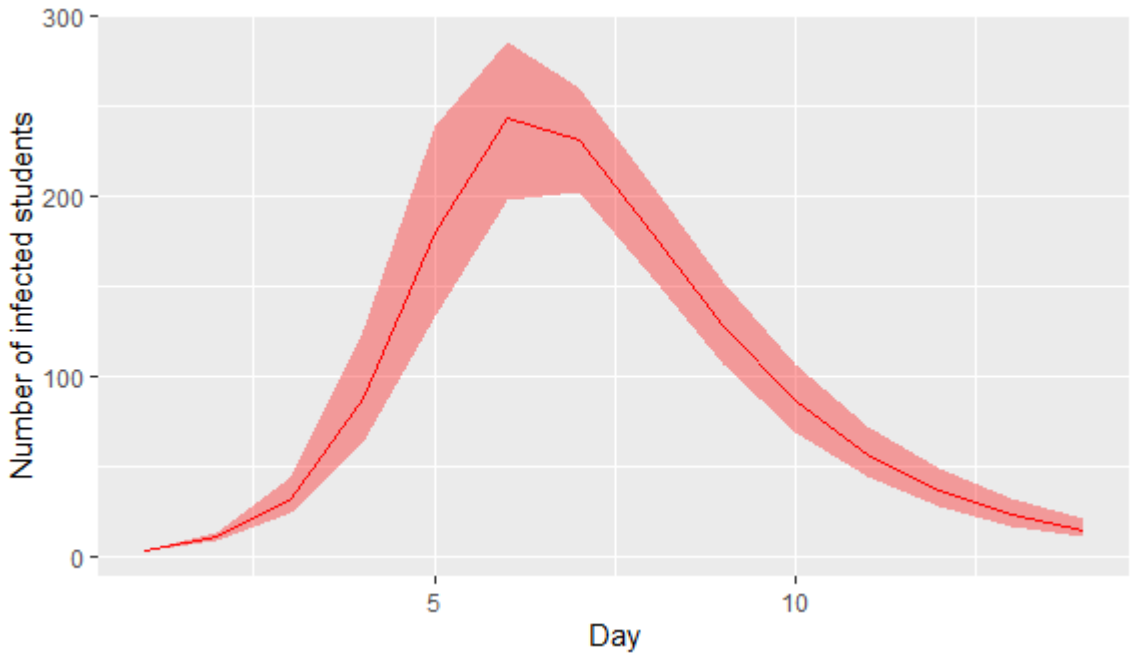
When we run it, we get followings plot.



Further, if we want to calculate not only the number students hospitalized but infected people at each time, we do followings,

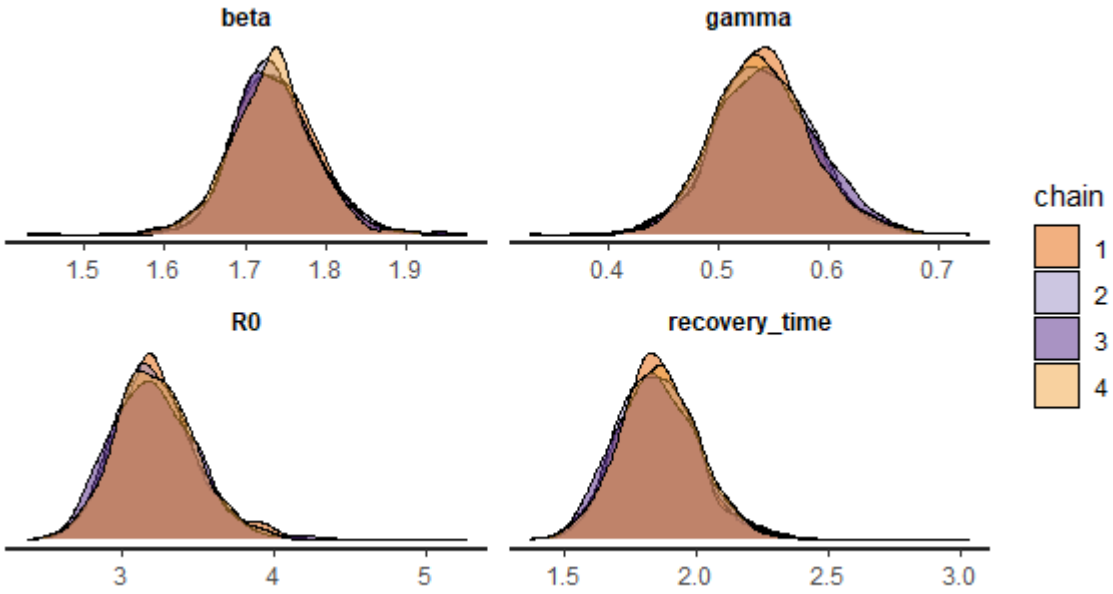
```
params <- lapply(t, function(i){sprintf("y[%s,2]", i)}) #number of infected for each day
smr_y <- as.data.frame(summary(fit_sir_negbin,
                               pars = params, probs = c(0.05, 0.5, 0.95))$summary)
colnames(smr_y) <- make.names(colnames(smr_y)) # to remove % in the col names

ggplot(smr_y, mapping = aes(x = t)) +
  geom_ribbon(aes(ymin = X5., ymax = X95.), fill = c_posterior, alpha = 0.35) +
  geom_line(mapping = aes(x = t, y = X50.), color = c_posterior) +
  labs(x = "Day", y = "Number of infected students")
```

When we run it, we get the following.



Results and Further Research:

The plots we see above, and inference table provide us the information we need as to predict how the disease is spreading among populations.



In this figure, 4 different chains shown by four different colours are almost in complete agreement with one another and thus giving us reliable estimates, we need. The average values as we see in bell shaped curves give us the parameters that determine the whole transmission dynamics of the disease. As I have mentioned above R_0 being one of the most significant parameters is found to be more than 3. This indicates that the influenza that spread in the boarding school falls into a severe epidemic. In order for this epidemic to have under control, they needed to bring down R_0 to at least 1 or below by introducing social distancing measures or some other effective measures such as vaccination.

Further we also checked whether our model captures the structures of the data and got followings plot.

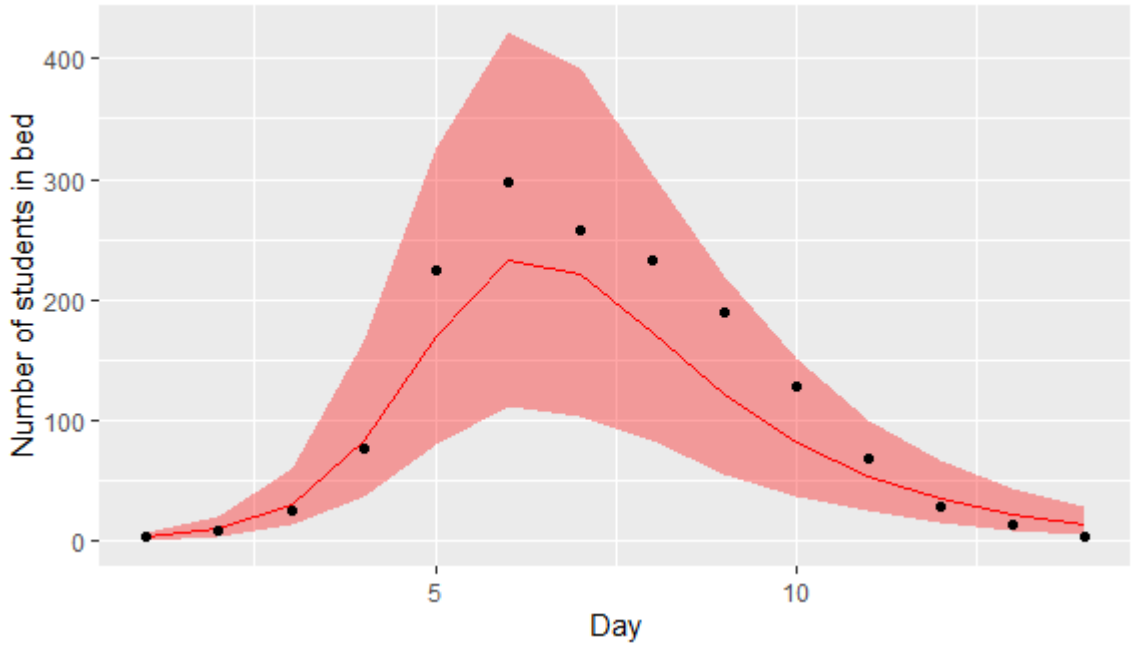


Fig: Best fit curve for number of hospitalized students

By following Bayesian statistics, we sampled $Y_{prediction}$ from $p(Y_{prediction}|Y)$ and plotted a best fitted curve for students in bed as above at 90% level of confidence meaning 10% observed data is assumed to fall outside of this interval. This method, oftentimes, called posterior predictive check, helped us confirm that our model is indeed capturing the structure of the data. We observed in above plot that our model is best fitted to the data and includes the variation within data as well.

Moreover, we did plot the same, for number of infected people at each time

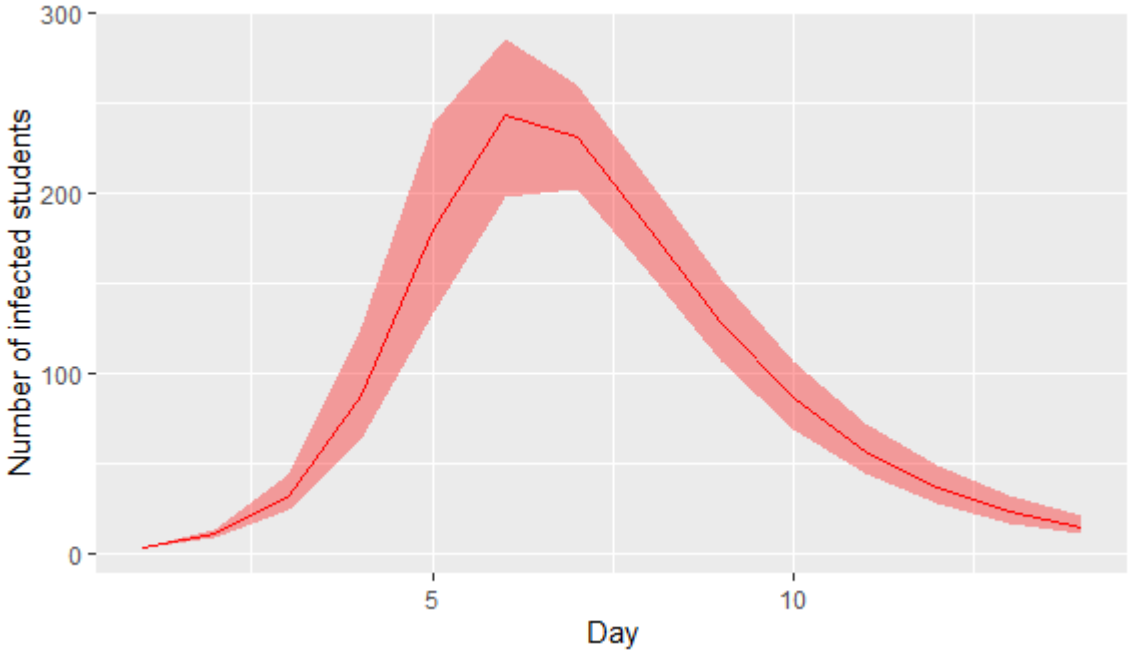


Fig: Best fit curve for true number of infected individuals at each point of time

Here also we saw that our model is best fitted with the observed data confirming us the inference we drew is reliable and valid. These plots, in compartmental modelling give us key insights into the dynamics of the disease. With the help of these estimates, public health officials can come up with intervene measures as to how to rein in the spread of the disease and ultimately beat it. The same techniques apply for most of the infectious diseases in compartmental modelling. Depending upon the complexities of the model involved, we try to tweak the models by adding more compartments. While adding more compartments may increase the accuracy of the predictions, it can also be computationally heavy and requires demanding algorithmic techniques. However, MCMC is used in most of such predictive techniques and for COVID-19 disease as well.

To recapitulate, I would like to conclude by saying that our main objective was to use real COVID-19 data and use it in SIR model to estimate the same model parameters. Therefore, I tried not to digress from COVID to other infectious diseases. However, in our final days, my supervisor and I found that the inference we got are not reliable and presentable as evidenced by significant variation in $\hat{R}$ value implying that our Markov chains are asynchronous to one another. The reason for this maybe the data we intended to use does not fit SIR model. The other reason maybe our Markov chain could not run simply because sample being too small from the huge population size of Norway (5.3m) and therefore maybe requiring stronger computers to run Markov chains. Nonetheless, we have tried and showed the same techniques we may have otherwise used for COVID-19 disease. And by doing so, we hope we have achieved the main objective of my thesis.

i) R codes for ggplot 6: SIR model.

Source (http://epirecip.es/epicookbook/chapters/sir/r_desolve)

```
library(deSolve)
library(reshape2)
sir_ode <- function(times,init,parms){
  with(as.list(c(parms,init)), {
    # ODEs
    dS <- -beta*S*I
    dI <- beta*S*I-gamma*I
    dR <- gamma*I
    list(c(dS,dI,dR))
  })
}
parms <- c(beta=0.1,gamma=0.05)
init <- c(S=0.99,I=0.01,R=0)
times <- seq(0,200,length.out=2001)
sir_out <- lsoda(init,times,sir_ode,parms)
sir_out_long <- melt(as.data.frame(sir_out),"time")
ggplot(sir_out_long,aes(x=time,y=value,colour=variable,group=variable))+
  # Add line
  geom_line(lwd=2)+
```

```
  #Add labels
  xlab("Time")+ylab("Number")
```

ii) R codes for ggplot7: SEIR Model.

Source : (http://epirecip.es/epicookbook/chapters/seir/r_desolve)

```
# Load deSolve library
library(deSolve)
# Function to return derivatives of SEIR model
seir_ode<-function(t,Y,par){
  S<-Y[1]
  E<-Y[2]
  I<-Y[3]
  R<-Y[4]

  beta<-par[1]
  sigma<-par[2]
  gamma<-par[3]
  mu<-par[4]

  dYdt<-vector(length=3)
  dYdt[1]=mu-beta*I*S-mu*S
```

```
  dYdt[2]=beta*I*S-(sigma+mu)*E
  dYdt[3]=sigma*E-(gamma+mu)*I

  return(list(dYdt))
}
# Set parameter values
beta<-520/365;
sigma<-1/60;
gamma<-1/30;
mu<-774835/(65640000*365) # UK birth and population figures 2016
init<-c(0.8,0.1,0.1)
t<-seq(0,365)
par<-c(beta,sigma,gamma,mu)
# Solve system using lsoda
sol<-lsoda(init,t,seir_ode,par)
```

REFERENCES:

Alimohamadi, Y., Taghdir, M., & Sepandi, M. (2020). Estimate of the basic reproduction number for COVID-19: a systematic review and meta-analysis. *Journal of Preventive Medicine and Public Health, 53*(3), 151.

Altamimi, A., Ahmed, A. E. J. J. o. I., & Health, P. (2020). Climate factors and incidence of Middle East respiratory syndrome coronavirus. *13*(5), 704-708.

Anderson, R. M., & May, R. M. J. N. (1985). Vaccination and herd immunity to infectious diseases. *318*(6044), 323-329.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. J. M. l. (2003). An introduction to MCMC for machine learning. *50*(1), 5-43.

Beichl, I., Sullivan, F. J. C. i. S., & Engineering. (2000). The metropolis algorithm. *2*(1), 65-69.

Biggerstaff, M., Cauchemez, S., Reed, C., Gambhir, M., & Finelli, L. J. B. i. d. (2014). Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature. *14*(1), 1-20.

Bolstad, W. M., & Curran, J. M. (2016). *Introduction to Bayesian statistics*: John Wiley & Sons.

Brauer, F., & Castillo-Chávez, C. (2001). Basic ideas of mathematical epidemiology. In *Mathematical Models in Population Biology and Epidemiology* (pp. 275-337): Springer.

Britton, T., Ball, F., & Trapman, P. J. S. (2020). A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *369*(6505), 846-849.

Chakhunashvili, G., Wagner, A. L., Power, L. E., Janusz, C. B., Machablishvili, A., Karseladze, I., . . . Gray, G. C. J. P. o. (2018). Severe Acute Respiratory Infection (SARI) sentinel surveillance in the country of Georgia, 2015-2017. *13*(7), e0201497.

Chen, Y.-C., Lu, P.-E., & Chang, C.-S. (2020). A Time-dependent SIR model for COVID-19. *Transactions on Network Science and Engineering*.

Dietz, K. (1993). The estimation of the basic reproduction number for infectious diseases. *Statistical methods in medical research*

2(1), 23-41.

Duong, D. (2021). What's important to know about the new COVID-19 variants? In: Can Med Assoc.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*: CRC press.

Grinsztajn, L., Semenova, E., Margossian, C. C., & Riou, J. J. a. p. a. (2020). Bayesian workflow for disease transmission modeling in Stan.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review, 42*(4), 599-653.

Kendall, D. G. (2020). Deterministic and stochastic epidemics in closed populations. In *Contributions to Biology and Problems of Health* (pp. 149-166): University of California Press.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical physical character, 115*(772), 700-721.

Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H., & Lipsitch, M. (2020a). Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period, 368*(6493), 860-868.

Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H., & Lipsitch, M. J. S. (2020b). Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *368*(6493), 860-868.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., . . . Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*

172(9), 577-582.

Liu, X., Huang, J., Li, C., Zhao, Y., Wang, D., Huang, Z., & Yang, K. J. E. r. (2021). The role of seasonality in the spread of COVID-19 pandemic. *195*, 110874.

Mahase, E. (2021). Covid-19: What new variants are emerging and how are they being investigated? In: British Medical Journal Publishing Group.

Monto, A. S., DeJonge, P. M., Callear, A. P., Bazzi, L. A., Capriola, S. B., Malosh, R. E., . . . Petrie, J. G. (2020). Coronavirus Occurrence and Transmission Over 8 Years in the HIVE Cohort of Households in Michigan. *The Journal of Infectious Diseases, 222*(1), 9-16. doi:10.1093/infdis/jiaa161 %J The Journal of Infectious Diseases

Pandey, G., Chaudhary, P., Gupta, R., & Pal, S. J. a. p. a. (2020). SEIR and Regression Model based COVID-19 outbreak predictions in India. *SEIR and Regression Model based COVID-19 outbreak predictions in India*.

Pinsky, M., & Karlin, S. (2010). *An introduction to stochastic modeling*: Academic press.

Randolph, H. E., & Barreiro, L. B. J. I. (2020). Herd immunity: understanding COVID-19. *52*(5), 737-741.

Rushton, S., & Mautner, A. J. B. (1955). The deterministic model of a simple epidemic for more than one community. *The deterministic model of a simple epidemic for more than one community*, *42*(1/2), 126-132.

Sajadi, M. M., Habibzadeh, P., Vintzileos, A., Shokouhi, S., Miralles-Wilhelm, F., & Amoroso, A. J. J. n. o. (2020). Temperature, humidity, and latitude analysis to estimate potential spread and seasonality of coronavirus disease 2019 (COVID-19). *3*(6), e2011834-e2011834.

Snow, J. (1936). Snow on cholera. London: Humphrey Milford. In: Oxford University Press.

Teles, P. J. a. p. a. (2020). A time-dependent SEIR model to analyse the evolution of the SARS-CoV-2 epidemic outbreak in Portugal.